Scenario Description

- WVCorp: the company you (the data scientist) work for
  - WVCorp has user forums and discussion boards for each of their products, where customers can discuss issues and features.
  - "Buzz": when a topic on the user forum has a very high activity level -- considered an indication of user interest in that topic.
- eRead: WVCorp's ebook reader product
- TimeWrangler: WVCorp's time-management app
- BookBits: A competitor's ebook reader product
- GCal: a third-party cloud-based calendar infrastructure that TimeWrangler can integrate with

Notes to describe the fictional world this presentation takes place in

Predicting Buzz

WVCorp Data Science Team

Notional Project Sponsor Presentation

If Only We'd Known...

<u>eRead vs. BookBits</u>

- eRead: Best selling indie ebook reader - until BookBits v.2
- Hot new BookBits feature: shared bookshelves
    - Our one-at-a-time book lending didn't compete
- Estimated <u>$25M lost revenue</u> on product sales

Here we establish the business problem that motivates this project: WVCorp is not in touch with customer needs.

eRead is WVCorp's product (an ebook reader). BookBits is the primary competing product in that space.

Could we Have Caught This?

- eRead Forum discussions:
  - Sharing a booklist with a friend, to grab from as they pleased
  - Sharing a book with a group of friends (first-come-first-serve)
- Whenever these questions arose, the discussion was lively
  - Suggestions, work-arounds, kludges, "me too"s
  - A shared bookshelf (like BookBits) would have met these recurring needs
- There was **Buzz** around this issue! But we ignored it. Or didn't find it.
  - Labor intensive to continually keep up with forum activity

Here we establish that the problem can potentially be solved, if we can exploit the information available in the product forums. But in the current workflow, monitoring the forums is a pain point. In a real presentation, you might use a screenshot of a relevant forum discussion.

Goal: Catch it Early

- Predict which topics on our product forums will have persistent buzz
  - Features customers want
  - Existing features users have trouble with
  - Persistent buzz, not ephemeral or trendy issues
    - Persistence = real, ongoing customer need

Here we explain the project goal, in the context of the existing business need. In a real presentation, perhaps another appropriate screenshot of a forum discussion.

Pilot Study

- Collected three weeks of data from forum
- Trained model on Week 1 to identify which topics will buzz in Weeks 2/3
  - Buzz = Sustained increase of 500+ active discussions in topic/day, relative to Week 1, Day 1
  - Compared predicted results to topics that actually buzzed
  - Feedback from team of five product managers -- how useful were the results?

High level, non-technical description of what we did

Results

- Reduced manual scan of forums by over a factor of 4
  - Scan 184 topics -- not 791!
  - PMs: 75% of identified topics produced "valuable insight"
- Found 84% of about-to-buzz topics
- Low (20%) false positive rate

	Predicted No Buzz	Predicted Buzz	
No Buzz	579	35	614
Buzz	28	149	177
Total	607	184	791

topics predicted to buzz that didn't

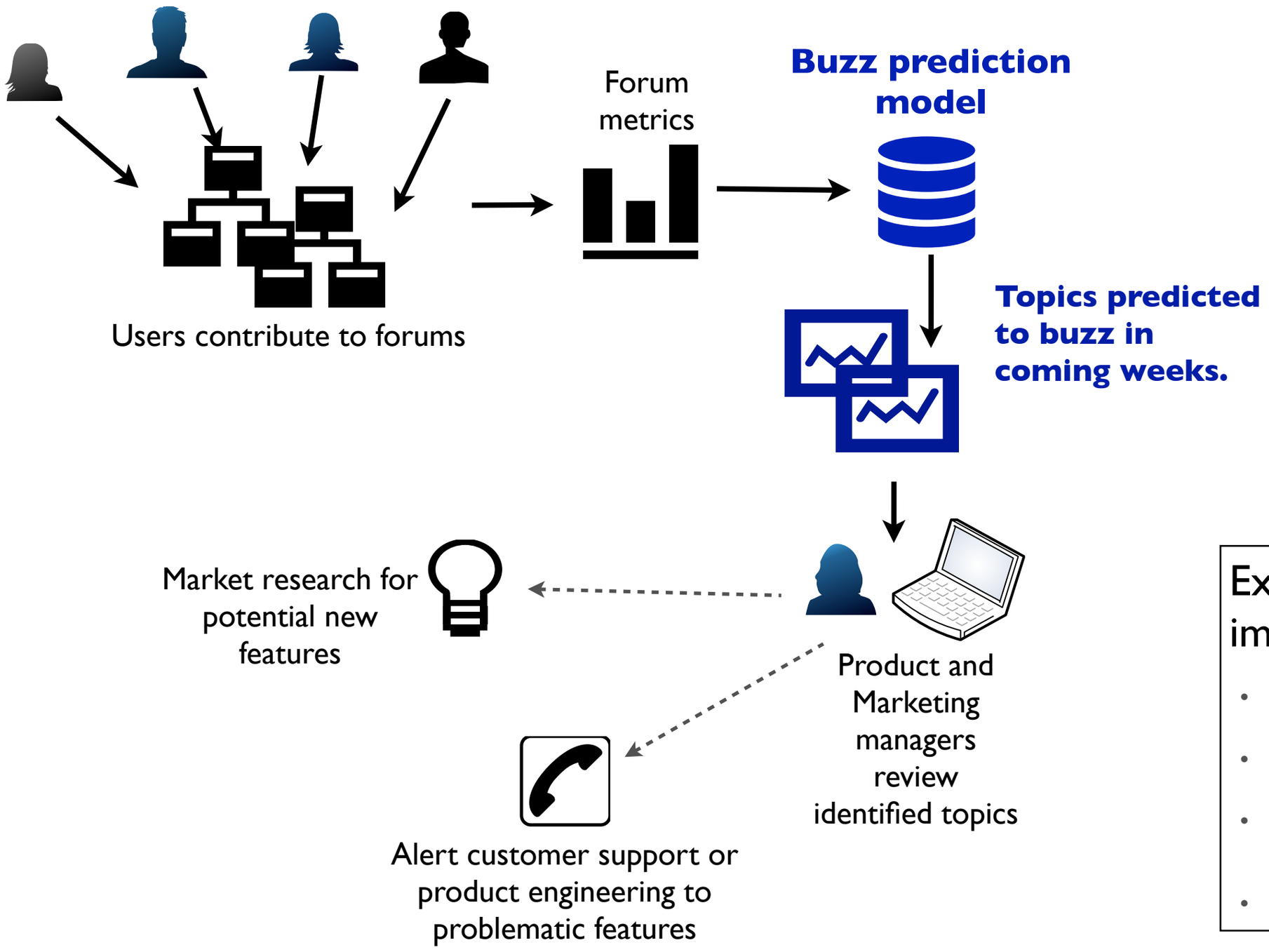
about-to-buzz topics that were missed

topics the PMs can skip

topics the PMs have to review

State the outcome of the project, how and why it is successful. Keep it concrete and non-technical. Don't talk about precision/recall, but rather about how much the workload for Product/Marketing managers is reduced, and how many potentially useful results the model found, and how many it missed.

How it Works



Exploit already implemented metrics

- # Authors/topic
- # Discussions/topic
- # Displays of topic to forum users
- etc.

A high-level, not-too-technical explanation of what we did, and where it fits.

Buzz Model

- Random Forest Model
  - Many "experts" voting
  - Runs efficiently on large data
  - Handles a large number of input variables
  - Few prior assumptions about how variables interact, or which are most relevant
  - Very accurate

Usually, the audience for this presentation is not terribly interested in the modeling methodology used. One high level slide should be sufficient, focusing on why you picked it (why it's the 'right' algorithm). Be prepared to skim over this slide quickly, or even skip it.

Example: Catching An Issue Early

- Topic: TimeWrangler →GCal Integration
  - # discussions up since GCal v. 7 release
    - GCal events not consistently showing up; mislabeled.
    - TimeWrangler tasks going to wrong GCalendar
  - **Hot on forums before hot in customer support logs**
    - Forum activity triggered the model two days after GCal update
    - Customer support didn't notice for a week

We give an example of an interesting finding from the project: an issue that we identified faster through the forums than through customer support logs. TimeWrangler is another WVCorp product, for time management; GCal is a popular third-party cloud-based calendar service. In a real presentation, you'll probably have more than one example.

Next Steps

- Further reduce PM workload, give them better customer intelligence.
  - New metrics for better prediction
    - Record if discussion activity is growing/shrinking, and how fast
    - Why do new forum users join? What question did they come to ask?
    - Goal: Find 98% of impending buzz, 10% false positive rate
  - Efficiently route buzz info to relevant Product Managers, Marketing, and Customer Support groups

Discuss how the results of the project can be improved, make recommendations for next steps.

Thank You